Walsh変換の集団遺伝学への応用

AUTHOR(S):

松本, 啓史

数理解析研究所講究録
第 870 巻 1994 年 32-42

Walsh 変換の集団遺伝学への応用

松本啓史

東京大学工学部計数工学科修士一年

1 背景

Walsh 変換は当初 Fourier 変換のデジタル版とし作られたが、1980 年 Bethke [2] が初めてそれをエピスタシスを表すのに用いた。また、それよりはやく 1980 年には Notohara,Matuda,Ishii [3] が 2 ビットの場合の Walsh 変換と同じものを別途に考案し 2 座位の場合に用いている。しかしながら、前者は系の時間発展を扱っていないし、後者は時間発展は扱っているが座位の数は限られている。

本稿は、 Notohara,Matsuda,Ishii[3] を任意座位数に拡張しようとする試みである。

2 扱うモデル

1. 2-allel、 l-loci の場合を考える。染色体を長さ l のビット列で表す。 長さ l のビット列とは、0,1 を l 個並べたものである。

$$i = i_1 i_2 ... i_l \qquad (i_p = 1\, or\, 0)$$

   今後、適宜ビット列という言葉と染色体という言葉とを同義に使う。

2. 一倍体

3. 連続時間

4. 遺伝的浮動は無視

5. 第 p 座位での突然変異において、 $1 \rightarrow 0, 0 \rightarrow 1$ の起こる確率は同じで、 μ_p である。

6. マルサス径数は時間によらない。染色体 i のマルサス径数を m_i とあらわす。

また、染色体 i の集団の中に占める割合を x_i と置く。$[\dot{x}_i]_{sel}, [\dot{x}_i]_{rec}, [\dot{x}_i]_{mut}$ はそれぞれ x_i の自然選択による変化、交叉、突然変異による変化の事である。

上記の仮定により、

$$[\dot{x}_i]_{sel} = x_i(m_i - \overline{m}) \tag{1}$$

となる。ただし、

$$\overline{m} = \sum_i m_i x_i$$

である。また、

$$[\dot{x}_i]_{rec} = -\frac{1}{2}\sum_S \gamma_S(x_i - x_{i_S} x_{i_{\overline{S}}}) \tag{2}$$

(2) 式中に現れる記号の定義は以下の如くである。

1. $S = \{s_1, s_2, ..., s_k | s_p = 1, 2, ...l, s_p \neq s_q, (p \neq q)\}$

   つまり S はいくつかの座位を集めた集合である。座位をいくつか集めた集合を ブロックという。

2. $\bar{S} : S$ の補集合。

3. $i_S = i_{s_1} i_{s_2} ... i_{s_l}$

   すなわち i の中から第 s_1 第 s_2... 第 s_k ビットを抜き出してきたビット列。

4. $$x_S = \sum_{j: j_S = i_S} x_j$$

   和は第 s_1 第 s_2... 第 s_k ビットが i と一致するビット列をすべて足す。

5. $\gamma_S : s_p \in S$ なる第 s_p 座位をすべて入れ換える交差の 起こる確率。 あきらかに $\gamma_S = \gamma_{\overline{S}}$。

(2) 式で、和の前に $\frac{1}{2}$ が掛かっているのは、 S を入れ換える交叉と $\overline{S}$ を入れ換える交叉は実は同じものだから、すべての S について和をとると、同じ項を 2 回足してしまうからである。

以後、i_S の一致しているビット列全体を集めた集合を S 上のスキーマタ、という。例えば $l=2$ で $S=\{1\}$ ならば、{10,11},{00,01} はともに S 上のスキーマタである。そして {10,11} を 1*, {00,01} を 0* と表す。* はここのビットには何が来ても良い、という印である。その他に、

$$\begin{aligned} 1*0 &= \{100,110\} \\ 0** &= \{000,001,010,011\} \end{aligned}$$

などのように使う。また、

$$x(1*1)=x_{101}+x_{111}$$

などと定義する。なお、$i=101, j=111, S=\{1,3\}$ とすると $1*1$ は S 上のスキーマタで i、j を含み、$x(1*1)=x_{i_S}=x_{j_S}$ がなり立つ。

今定義した記号を用いると、突然変異による時間変化は

$$[\dot{x}(**\ldots*\overset{\overset{p}{\vee}}{0}*\ldots*)]_{mut}=-\mu_p(x(**\ldots\overset{\overset{p}{\vee}}{0}**\ldots*)-x(**\ldots*\overset{\overset{p}{\vee}}{1}*\ldots*)) \tag{3}$$

と書ける。

3 エピスタシスについて

$\varphi(i)$ を実数値をとるビット列の関数とし、$\varphi_S(i)$ をビット列の関数で $s_p\in S$ なる第 s_p ビットのみの関数になっているものとしよう。関数 $\varphi(i)$ を次の様に展開する [1]。

$$\varphi(i)=\sum_S\varphi_S(i)=\varphi_0+\sum_{S:|S|=1}\varphi_S(i)+\cdots+\sum_{S:|S|=l}\varphi_S(i) \tag{4}$$

なお、φ_0 はある定数で $S=\phi$ の時の $\varphi_S(i)$ であり、$|S|$ は S の要素数である。各 $\varphi_S(i)$ は以下のようにして決められる。まず φ_0 を $\sum_i(\varphi(i)-\varphi_0)^2$ を最小にするように決め、次に $|S|=1$ の $\varphi_S(i)$ を $\sum_i(\varphi(i)-\varphi_0-\sum_{S,|S|=1}\varphi_S(i))^2$ を最小にするように決め、... 以下同様に $|S|=l-1$ の $\varphi_S(i)$ 迄決める。$|S|=l$ の $\varphi_S(i)$ は実は一つしかないから $\varphi_S(i)=\varphi(i)-\sum_{S,|S|<l}\varphi_S(i)$ できめる。

この決め方から、$|S|=l$ の $\varphi_S(i)$ を各ビット独自の $\varphi(i)$ への寄与、$|S|=2$ の $\varphi_S(i)$ を二つのビット間の相互作用、$|S|=3$ の $\varphi_S(i)$ を 3 ビット間相互作用、... といって良いだろう。従って、$\varphi(i)$ が m_i の時 $|S|\geq 2$ の $\varphi_S(i)$ はエピスタシスを表すことになる。

この $\varphi_S(i)$ を簡単に計算する方法を Walsh 変換が与える。

4 Walsh 関数

長さ 1 のビット列に対する walsh 変換は次のように定義する。

$$[i_p, i_p] = \begin{cases} -1 & i_p = j_p = 1 \\ 1 & otherwise \end{cases} \tag{5}$$

そして、長さ l のビット列に対する walsh 変換を次のように定義する。

$$[i,j] = \prod_{p=1}^{l} [i_p, j_p] \qquad \left. \begin{array}{l} i = i_1 i_2 \ldots i_l \\ j = j_1 j_2 \ldots j_l \end{array} \right\} \text{ビット列} \tag{6}$$

定義にしたがった簡単な計算で、次の二つの定理が導ける。

定理 4.1 *1.*

$$[i,j] = [j,i]$$

2.

$$[i,0] = \prod_{p=1}^{l} [i_p, 0] = 1$$

ただし 0 は全てのビットが 0 のビット列をも表す。（以後も同様。）

3.

$$\sum_i [i,j] = 2^l \delta_{j,0}$$

4.

$$[i,j] = [i_S, j_S][i_{\bar{S}}, j_{\bar{S}}]$$

定理 4.2

$$\frac{1}{2^l} \sum_j [i,j][j,k] = \delta_{i,j} \tag{7}$$

ただし、i, j, k は長さ l のビット列。

定理 4.2より、Walsh 関数系は直交関数系である。

5 Walsh 変換

$\varphi(i)$ の Walsh 変換 $\Phi(j)$ とは、

$$\Phi(i) = \frac{1}{2^l}\sum_i [i,j]\varphi(i) \tag{8}$$

で定義される。すると、定理 4.2から、次の定理が容易に証明できる。

定理 5.1

$$\varphi(i) = \sum_i [i,j]\Phi(i) \tag{9}$$

なお、定理 5.1で現れる $\sum_i [i,j]\Phi(i)$ を $\Phi(i)$ の Walsh 逆変換という。 Walsh 逆変換と Walsh 変換とでは、定数 $\frac{1}{2^l}$ だけ違う。

今、 $S = \{s_1, s_2, ...s_k | s_p = 1, 2, ...l\}$ としたとき、

$$1(S) = 0...0 \overset{\overset{s_1}{\vee}}{1} 0...0 \overset{\overset{s_2}{\vee}}{1} 0...0 \overset{\overset{s_k}{\vee}}{1} 0...0$$

とする。また、1({p})、1({p,q}) などをはしょって 1(p)、1(p,q) と書く。

定理 5.2 $\varphi_S(i)$ と $\Phi(i)$ の間には次の関係がある [4]。

$$\varphi_S(i) = [i, 1(S)]\Phi(i) \tag{10}$$

定理 5.3 h をある S 上のスキーマタとし、$\varphi(h) = \frac{1}{|h|}\sum_{i\in h}\varphi(i)$ と定義する。($|h|$ はスキーマタ h に含まれる要素の数。) $\varphi(h)$ は $\varphi(i)$ におけるブロック S の寄与を表す。$\varphi(h)$ は次の式で表せる [2]。

$$\varphi(h) = \sum_{h:j_S}[h,j]\Phi(j) \tag{11}$$

なお、h は h の要素で $\bar{S}$ に属するビットは 0 であるビット列をも表す。

証明は両方とも初等的な計算による。x_i, m_i の Walsh 変換をそれぞれ X_j, M_j とする。定理 5.2より M_j はの $|S| = 1$ とき各座位独自の効果で $|S| \geq 2$ の時エピスタシスを表す。

6 Z 座標の定義と意味

$$
\begin{aligned}
z_i &= ln\, x_i & (12)\\
Z_j &= \frac{1}{2^l}\sum_i [i,j] z_i & (13)\\
\mathbf{S} &= \{S_1, S_2, ...S_n | \bigcup_{p=1}^{n} S_p = \{1,2,...,l\}, S_p \cap S_q = \phi, (p \neq q)\} & (14)\\
Z_j^{\mathbf{S}} &= \frac{1}{2^l}\sum_i [i,j]\, ln \frac{x_i}{x_{1_{S_1}} x_{2_{S_1}} \cdots x_{n_{S_1}}} = \frac{1}{2^l}\sum_i [i,j]\, ln \frac{x_i}{\prod_{S\in\mathbf{S}} x_{i_S}} & (15)
\end{aligned}
$$

と定義する。すぐに分かるように、

$$x_i = x_{1_{S_1}} x_{2_{S_1}} \cdots x_{n_{S_1}} = \prod_{S\in\mathbf{S}} x_{i_S} \Rightarrow Z_j^{\mathbf{S}} = 0 \tag{16}$$

であり、また、

$$Z_j^{\mathbf{S}} = Z_j - \frac{1}{2^l}\sum_{S\in\mathbf{S}} \Big(\sum_{i_{\bar{S}}} [i_{\bar{S}}, j_{\bar{S}}]\Big)\Big(\sum_{i_S} [i_S, j_S]\, ln\, x_{i_S}\Big) \tag{17}$$

である。さらに定理4.1より、

$$\sum_{i_{\bar{S}}} [i_{\bar{S}}, j_{\bar{S}}] = 2^{|\bar{S}|}\delta_{0,j_{\bar{S}}} \tag{18}$$

が成り立つ。ここである $S,T \in \mathbf{S}$ について $j_S \neq 0, j_T \neq 0$ を満たすようなビット列 j を持ってこよう。すると、どんな $S \in \mathbf{S}$ についても $j_{\bar{S}}$ となる。従って (17) 式から、こうした j をとってくれば、

$$Z_j^{\mathbf{S}} = Z_j \tag{19}$$

となっている。ゆえに、

$$I_{\mathbf{S}} = \{j | \exists_{S,T\in\mathbf{S}}\ j_S \neq 0, j_T \neq 0\}$$

とすると (16) 式,(19) 式から、

$$x_i = \prod_{S\in\mathbf{S}} x_{i_S} \Rightarrow Z_j = 0 \tag{20}$$

が決論出来る。このことから、Z_j は連鎖平衡からのずれを表す指標になる。

7 Y 座標

Y 座標は

$$Y_j = 2^l X_j = \sum_i [i,j] x_i$$

と定義される (x_i の逆 Walsh 変換)。定理 5.3から h を S 上のスキーマタとすると、

$$x(h) = \frac{1}{2^{|S|}} \sum_{j:j_{\bar{S}}=0} [h,j] Y_j \tag{21}$$

となる。$Y_0 = \sum_i x_= 1$ であるから、たとえば

$$\begin{aligned} x(0**) &= \frac{1}{2}(1+Y_{100}) && (22)\\ x(1**) &= \frac{1}{2}(1-Y_{100}) && (23)\\ & && (24) \end{aligned}$$

などとなる。つまり h を S 上のスキーマタとすると、$x(h)$ は $Y_j(j_{\bar{S}}=0)$ だけで計算できる。

8 YZ 混合座標

$$J_{\mathrm{S}} = \{j | \forall_{S \in \mathrm{S}}\, j_S \neq 0 \Rightarrow j_{\bar{S}} = 0\} \tag{25}$$

と定義する。そして

$$Z_j\ j \in I_{\mathrm{S}} \ \ ,\ \ Y_j\ j \in J_{\mathrm{S}} \tag{26}$$

を座標にとると、

$$I_{\mathrm{S}} \cup J_{\mathrm{S}} \cup \{0\} = \text{長さ}l\text{のビット列全体} \tag{27}$$

なので、座標変数の数は十分にあり、この座標で系の状態を少なくとも局地的には定め得る。これを YZ 混合座標系という。先の節の結果から、ブロック S の周辺分布 x_{i_S} は $Y_j\,(j \in J_{\mathrm{S}})$ で計算できる。また、6 節の議論から $Z_j\,(j \in I_{\mathrm{S}})$ は x_i の $\prod_{S \in \mathrm{S}} x_{i_S}$ からのずれをあらわしている。

9 Z 座標の時間発展 (1)

$$[\dot{Z}_j]_{sel} = \begin{cases} M_j & j \neq 0 \\ M_j - \bar{m} & j = 0 \end{cases} \tag{28}$$

$$[\dot{Z}_j]_{rec} = -\frac{1}{2^{(l+1)}} \sum_S \gamma_S \sum_i [i,j](1 - \frac{x_{i_S} x_{i_{\bar{S}}}}{x_i}) \tag{29}$$

簡単な計算で上二つの式が導かれる。なお、本節と次節では突然変異は扱わない。いま、

$$[\dot{Z}_j]^S_{rec} \equiv -\frac{1}{2^l}\gamma_S \sum_i [i,j](1 - \frac{x_{i_S} x_{i_{\bar{S}}}}{x_i})$$

すなわち染色体のうち、ブロック S だけを入れ換える交叉による Z_j の変化を $[\dot{Z}_j]^S_{rec}$ とする。そして

$$Z_j^S = -\frac{1}{2^l} \sum_i [i,j] ln \frac{x_i}{x_{i_S} x_{i_{\bar{S}}}}$$

と定義すると

$$Z_j^S = -\frac{1}{2^l} \sum_i [i,j] ln\,(1 - (1 - \frac{x_{i_S} x_{i_{\bar{S}}}}{x_i}))$$

であるから、

$$x_i \approx x_{i_S} x_{i_{\bar{S}}}$$

がなり立つ時には

$$Z_j^S \approx -\frac{1}{2^l} \sum_i [i,j](1 - \frac{x_{i_S} x_{i_{\bar{S}}}}{x_i}) \tag{30}$$

となり、従って

$$[\dot{Z}_j]^S_{rec} \approx -\gamma_S Z_j^S \tag{31}$$

となる。またもし $j_S = 0$ かつ $j_{\bar{S}} = 0$ がなり立つなら定理 4.1 より

$$\sum_{i_S} [i_S, j_S] = \sum_{i_{\bar{S}}} [i_{\bar{S}}, j_{\bar{S}}] = 0$$

となるので、

$$\begin{aligned} Z_j^S &= -\frac{1}{2^l} \sum_j [i,j] ln\, x_i - \frac{1}{2^l} \sum_{i_S} \sum_{i_{\bar{S}}} [i_S, j_S][i_{\bar{S}}, j_{\bar{S}}](ln\, x_{i_S} + ln\, x_{i_{\bar{S}}}) && (32) \\ &= Z_j - \frac{1}{2^l}(\sum_{i_S} [i_S, j_S]) \sum_{i_{\bar{S}}} [i_{\bar{S}}, j_{\bar{S}}] ln\, x_{i_S} - \frac{1}{2^l}(\sum_{i_{\bar{S}}} [i_{\bar{S}}, j_{\bar{S}}]) \sum_{i_S} [i_S, j_S] ln\, x_{i_{\bar{S}}} && (33) \\ &= Z_j && (34) \end{aligned}$$

である。それゆえ

$$[\dot{Z}_j]^S_{rec} \approx \gamma_S Zj \quad (j_S = 0,\, j_{\bar{S}} = 0,\, x_i \approx x_{i_S} x_{i_{\bar{S}}}) \tag{35}$$

なので $j = 1$ の時には

$$[\dot{Z}_1]_{rec} = \frac{1}{2}\sum_S [\dot{Z}_1]^S_{rec} \tag{36}$$

$$\approx -(\sum_S \gamma_S) Z_1 \tag{37}$$

となり、$M_1 = 0$ なら (29) 式と併せて

$$\dot{Z}_1 \approx -(\sum_S \gamma_S) Z_1 \tag{38}$$

である。他だし1は全てのビットが1のビット列をも表す。よって Z_1 は0付近で安定だと分かる。

10　Z 座標の時間発展 (2)

$$[\dot{x}_i]_{rec} = -\frac{1}{2}\sum_S \gamma_S (x_i - x_{i_S} x_{i_{\bar{S}}})$$

であったが、この式中に現れる S を S にふくまれる S に限定する。また、

$$\forall_{j \in I_{\mathsf{S}}}\, M_j = 0 \tag{39}$$

と仮定する。これは具体的には次の状況に対応する。

1. 染色体が $S_1, S_2, ..., S_n$ といういくつかのブロックに分かれる。
2. 各ブロック間にはエピスタシスがない。
3. 同じブロックに属する座位の間にはほとんど組換えがない。
4. ブロックとブロックの間では、組換えが頻繁におこる。

ここで

$$I'_{\mathsf{S}} = \{ j \,|\, \forall_{S \in \mathsf{S}}\, j_S \neq 0 \} \tag{40}$$

と定義すれば、前の節と同じようにして $\forall_{j \in I'_{\mathsf{S}}}$ に対して Z_j が0付近で安定であることがいえる。

そもそも交叉とは座位を組換えて連鎖平衡に系を導くものだとすれば、次のように (2) 式を”近似”してみるのも一つの考えであろう。

$$[\dot{x}_i]_{rec} = -\gamma(x_i - \prod_{S\in\mathbf{S}} x_{i_S}) \tag{41}$$

すると (39) 式の仮定のもと Z_j $(\forall_{j\in I_{\mathbf{S}}})$ が 0 付近で安定であることが、先の節のような議論で分かる。YZ 混合座標で考えれば、これはすなわち

$$x_i = \prod_{S\in\mathbf{S}} x_{i_S} \tag{42}$$

という状態が安定である、といい換えても良いことが理解できる。(39)、(42) が満たされている時

$$[\dot{x}(h)]_{sel} = x(h)(\hat{m}(h) - \overline{m}) \tag{43}$$

がなりたっている [4]。ただし h は $S\in\mathbf{S}$ 上のスキーマタで、$\hat{m}(h)$ は

$$\hat{m}(h) = \frac{1}{2^{|S|}}\sum_{i\in h} m_i$$

と定義される。

11　Y 座標の時間発展

いま交叉が極めて強く、$x_i = \prod_{p=1}^{l} x_{i(p)}$ が常になり立っているとしよう。この仮定のもと簡単な計算で、

$$[\dot{Y}_{1(p)}]_{sel} = (1 - Y_{1(p)}^2)(M_{1(p)} + \sum_{S:\,S\ni p}(M_{1(S)}\prod_{q\in S:\,q\neq p} Y_{1(q)})) \tag{44}$$

となることが分かる [4]。（系の状態は $x_{i(p)}$ だけで決まる。(21) 式、(24) 式を考えると、$Y_{1(p)}$ だけで決まっている。）$l=3$ の例ではたとえば

$$[\dot{Y}_{100}]_{sel} = (1 - Y_{100}^2)(M_{100} + M_{110}Y_{010} + M_{101}Y_{001} + M_{111}Y_{010}Y_{001}) \tag{45}$$

などとなる。Y_{100} の増加率が Y_{010}、Y_{001} に依存する仕方が、エピスタシスによる様が分かる。

なお、上記の仮定に関係なく

$$[\dot{Y}_{1(S)}]_{mut} = -2(\sum_{p\in S}\mu_p)Y_{1(S)} \tag{46}$$

であることが簡単な計算で分かる。

12 結語

本稿の目的は Notohara,Ishii,Matsuda[3] を一般化することであった。しかし、10節でやったように交叉の式を" 近似" したのはあまりきれいとはいえない。また、突然変異も盛り込めなかった。これらの欠点は、文献 [1] のように微分幾何を援用することによって解決されるだろう。

また、Walsh 変換が 2allel の時にしか使えないのは残念である。Walsh 変換を拡張する必要があり、おそらく可能であろう。

Use of Walsh Transform in Population Genetics

Department of Mathematical Engineering and Information Physics

University of Tokyo, Tokyo 113, Japan

Abstract

Walsh transform is introduced as a convienient method for revealing the magnitude of epistasis. Also,Y-coordinaate and Z-coordinae are proposed,becose those are useful in analysis of diferential equations which appear in population genetics.

参考文献

[1] Akin.E, "The Geometry Of Population Genetics", Lecture Notes of Biomathematics, Springer-Verlag(1978)

[2] Bethke,A.D, "Genetic Algorithms As Function Optimizers", University Microfilms International(1980)

[3] Notohara,M.,Ishii,K.,Matsuda,H., "Use of Orthothogonal Transformation in Population Genetics", Journal of Mathematical Biology, vol 6,pp.235-248(1978)

[4] 松本 啓史 "遺伝的アルゴリズムの数理的考察 ", 卒業論文 (1993)